2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Hyperparameters tuning in regularized system identification with nonzero prior means

Håvard B. Bjørkøy, Damiano Varagnolo

Abstract— **We consider the problem of identifying linear time invariant systems using regularization schemes, and address the fact that generally the mean value of the corresponding parameter prior is set to zero. We thus consider the scenario where it is beneficial to use a prior with nonzero-mean, where this mean moreover depends on some hyperparameters. We show how to construct such priors and do hyperparameter tuning by marginal likelihood, and since a parameter dependent mean may slow down optimization, we also derive an efficient and stable way of treating them, leading to an overall scheme whose leading order numerical complexity is the same as in the case where the prior mean is zero. The proposed method thus allows including new types of external information in the prior, and we exemplify how this extension can improve the existing regularization techniques.**

I. Introduction

This article explores ways of making established methods from regularized system identification utilize more types of information via data fusion (making them *more* Bayesian). A consequence of this will sometimes be that the prior mean has a mean dependent on some hyperparameters, that should be tuned, and we develop a way of doing so more efficiently than the naive approach. We are motivated by the fact that often there exists expert knowledge (that goes beyond only parameter correlations) that is not utilized in data-driven modeling, which the Bayesian methodology can incorporate [2]. An example of such expert knowledge is a (simple) physics based model of the observed system. Similar ideas are presented in [11], [14], though the perspective on regularization and hyperparameters here shines a new light on the topic.

Classical system identification relies on adjusting the model complexity through selecting the order of the dynamical model [17], i.e. the number of parameters and the model structure, alternating between parameter estimation and evaluating each model order and structure. The classical methods come with problems such as identifiability, persistence of excitation and ill-posedness, that is circumvented when using a *weakly informative* prior on the model [22]. The weakly informative prior *regularizes* the model, in the sense that the posterior inference is put on a reasonable level [12] (reasonable considering the evidence *and* the prior knowledge the prior reflects). Then the model complexity selection can be made in a continuous fashion, via tuning the regularization parameters. The Bayesian interpretation of this regularization approach is well known [21], [22]. A specific regularization penalty is equivalent to setting a specific prior $p(\theta)$ on the model parameters in the parametric Bayesian framework. E.g., L2 regularization penalties (such as ridge regression) translate well into Gaussian priors.

The main benefit of the simplest explicit regularization method, ridge regression, is that it can yield a reasonable posterior via selecting the regularization parameter in an opportune way (more on this below). More informed kernel design allows for more informative priors, embedding information on stability, smoothness and frequency response of the system [19] (with e.g. DI, TC, DC kernels). This is briefly why regularized methods have succeeded so well, and gained much attention in recent years [22]. These regularization methods introduce prior knowledge only via parameter correlations, i.e. the second moment of the model parameters. In this article we pursue methods that extend this, that construct and utilize more informative priors, being able to embed more or different information about the system (when available), in that they also carry information on the first moment of the model parameters (denoted θ). The models are arguably then *more* Bayesian.

Embedding more and different information into the prior often yields models that are more robust [26]. However, such information is often incomplete or uncertain, raising a need for hyperparameters. Empirical Bayes approaches that tune hyperparameters in the prior are indeed meaningful given that the prior information is meaningfully encoded [10]. The empirical Bayes procedure in the present article of selecting point estimates for the hyperparameters η is well-known and standard in regularized system identification; maximizing the output likelihood over η, given the prior $p(\theta)$ only. Maximizing this by evaluating the marginal likelihood function directly has some known issues; it is prone to numerical inaccuracies and has large computational complexity. We adapt two effective algorithms that amend these issues to the case where the parameter prior can have a nonzero mean.

Hyperparameters can of course alternatively be tuned based on cross-validation, though this may be criticized for e.g. tending to undersmooth [13] or simply for being a trial-and-error approach on a discretized set. Optimizing rather the generalized cross-validation (GCV) error function is another alternative to marginal likelihood [20], though a version (particularly an efficient one) for nonzero prior means is unknown. GCV is furthermore only approximate for regularizers other than ridge.

Maximizing the marginal likelihood with a nonzero prior

All the authors are with the Department of Engineering Cybernetics at the Norwegian University of Science and Technology, Trondheim, Norway. Emails: { havard.bjorkoy, damiano.varagnolo } @ntnu.no. This project received funds from the Norwegian Research Council through the Knowledge Building project number 326802, ”Electrical Conditions in Submerged Arc Furnaces - Identification and Improvement”.

mean has briefly been mentioned in [5], [8], though not one that depends on hyperparameters, but rather a fixed one obtained from a baseline model. This difference is significant when it comes to efficient tuning of hyperparameters in the prior. It is briefly pointed out in [22] how the mean may depend on hyperparameters, though doing empirical Bayes efficiently in that scenario lacks. In Section III we argue and exemplify how regularized identification problems may end up with a prior mean that depends on hyperparameters.

This work extends the main findings of Chen and Ljung in [6], in its turn a significant improvement of [4]. This work is also based on the manuscript and the derivations of the so-called *Algorithm 2* in [6]. Evaluating the marginal likelihood can be costly and ill-conditioned when approached naively, and the *Algorithm 2* partly amends this issue, though only for priors $p(\theta)$ with zero mean (or equivalently, constant wrt. η). Further improvements to *Algorithm 2* can be made whenever the inverse of the prior covariance matrix is available (meaning that computing it does not require matrix inversion), as shown by [3] and their *Algorithm 1*. The main findings in our article is an efficient algorithm for including also a mean dependent on hyperparameters, and is thus a generalization of *Algorithm 1* (when available) and *Algorithm 2*, as the procedures are adapted to also the case of nonzero prior means, derived in Section IV.

The application of the new algorithm is illustrated by simulation examples in Section V. As an example, a prior on the DC gain of the system can be embedded, though it requires tuning of a prior with nonzero mean that depends on additional hyperparameters. This metric is a decent indicator of out-of-sample performance of the model using the posterior mean, and also a meaningful metric considering that this Bayesian approach seeks to include properties of the true system to estimate it better. The tests indicate that the proposed method is able to improve modeling not only in theory but also in practice.

Together, these examples illustrate how regularized system identification can combine more general sources of prior information, with automatic hyperparameter tuning, which can lead to models that better describe the system to be identified and/or yield more robust modeling in scenarios with poorer data quality or adaptive models.

II. Problem definition and statement of contributions

We consider the standard case of a parameter affine model structure (corresponding to many linear model structures)

$$y = X\theta + e\ , \tag{1}$$

for modeling a dynamical system from measured input-output data. The regressors $X \in \mathbb{R}^{n\times k}$ are assumed to be known and deterministic, the output $y \in \mathbb{R}^n$ to be thus linear in the (unknown) parameters $\theta \in \mathbb{R}^k$, the parameters θ to follow some prior distribution discussed in more details below. The overall goal is to estimate the posterior distribution of θ. The noise e is for simplicity assumed i.i.d. as $e \sim N(0, \sigma^2 I_n)$ with variance σ^2 assumed to be known or to have been estimated from the data, e.g., by means of some low-bias FIR or ARX models [23], before this posterior distribution estimation step.

As for the prior available information on θ, we consider a parameterized Gaussian prior of the form

$$p\big(\theta;\eta\big) \sim N\big(m(\eta), V(\eta)\big) \tag{2}$$

with the functions $m(\eta)$ and $V(\eta)$ structurally known, but whose hyperparameter vector η is unknown.

For computing the posterior, the derivations implicitly consider the fact that for a given point value of the hyperparameters $\hat{\eta}$ one may select the specific prior moments $\hat{m} = m(\hat{\eta})$ and $\hat{V} = V(\hat{\eta})$, which is a commonly applied simplification for such hierarchical models [18]. When observing some measured data $\{X, y\}$, assuming this dataset to follow (1) means assuming a specific likelihood. By applying the Bayes rule, this likelihood and the prior in (2) may be combined giving the posterior parameter distribution

$$p(\theta|y) \sim N\left(m^*, V^*\right) \tag{3a}$$

$$V^* = \left(\hat{V}^{-1} + X^T X/\sigma^2\right)^{-1} \tag{3b}$$

$$m^* = V^*\left(X^T y/\sigma^2 + \hat{V}^{-1}\hat{m}\right)\ . \tag{3c}$$

Generally marginalizing this distribution over θ may require some calculus efforts; though in this linear-Gaussian case with deterministic X the marginal distribution of y given the prior (2) follows algebraically as

$$p(y|\eta) \sim N\left(Xm(\eta), XV(\eta)X^T + \sigma^2 I\right)\ . \tag{4}$$

The hyperparameter maximizing the marginal likelihood (4) follows thus as

$$\begin{aligned}\eta^* &= \arg\max_{\eta} p(y|\eta) \\ &= \arg\min_{\eta}\, (y - Xm)^T\left(XVX^T + \sigma^2 I\right)^{-1}(y - Xm) \\ &\qquad + \log\det\left(XVX^T + \sigma^2 I_n\right)\ ,\end{aligned} \tag{5}$$

where for notational brevity we assume the dependence of m and V on η as tacit. If m did not depend on η, then one could eliminate Xm from (5) [5], and this marginal likelihood would be identical to that considered in [6], which is though not the case in this work.

As for numerically finding η^*, the simplest approach would be to optimize this cost function, but evaluating (5) as it is has leading order of complexity $O(n^3)$. This can be computationally demanding when n is large. Typically in system identification $n \gg k$ [6] (though in machine learning it is at times the other way around). In the outlined framework, our main goal is thus to improve this naive and expensive approach of evaluating the marginal likelihood. To the best of our knowledge considering the body of available literature, $m(\eta) = 0$ is the typical assumption, while having $m(\eta) \neq 0$ in (2) is a novel contribution. In the remainder of the paper we thus:

1) motivate in Section III why the case $m(\eta) \neq 0$ is of practical relevance;

2) derive in Section IV an algebraically equivalent implementation to (5) with favorable numerical properties;
3) assess in Section V the performance of the proposed algorithm from numerical perspectives.

III. Motivating the case $m(\eta) \neq 0$

We now motivate the practical relevance of the case $m(\eta) \neq 0$. It may arise when fusing different apriori information items on θ into a single distribution $p(\theta)$. Multiple information items (in addition to measurements) are often available for physical modeling [25], and can benefit from being treated as uncertain, as acknowledged by Bayesian methods. In a sense, the following example is about making flexible gray-box models, with a Bayesian approach, making the real examples of application numerous [27]. We though keep the example generic for clarity and brevity.

Assume that two different sources of prior information on θ (e.g., by means of efforts translating some expert knowledge into parametric forms as in [15]) are encoded in two distributions along some subspace of $\mathbb{R}^k$, i.e., assume that θ follows

$$p_1\big(A\theta\big) \sim N\big(w, \Sigma_1(\eta)\big) \qquad p_2\big(B\theta\big) \sim N\big(0, \Sigma_2(\eta)\big) \quad (6)$$

where $(A\theta) \in \mathbb{R}^a$, $(B\theta) \in \mathbb{R}^b$, A and B are deterministic, $w \in \mathbb{R}^a$, and $\Sigma_1(\eta)$ & $\Sigma_2(\eta)$ are positive definite matrices for any hyperparameter vector η of practical meaning (that implicitly means that throughout the article we assume that admissible values for η are only those that render any covariance matrices positive definite).

Independently on whether w is fully known (fixed) or also parametrized in η, we show now that fusing the two priors p_1 and p_2 into one leads to a novel prior whose mean is known only partially, i.e. depends on η.

Consider then either one of the two following situations: *1)* $A\theta$ and $B\theta$ are jointly Gaussian with a covariance Σ_{12} either fully known or known through $\Sigma_{12} = \Sigma_{12}(\eta)$, or *2)* there is no available information on the joint distribution of $A\theta$ and $B\theta$.

As for case *1)*, the assumptions yield that

$$p\left(\begin{pmatrix} A \\ B \end{pmatrix}\theta\right) \sim N\left(\begin{pmatrix} w \\ 0 \end{pmatrix}, \Sigma(\eta)\right) ,$$
$$\Sigma(\eta) := \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_2 \end{pmatrix} .$$

Defining the precision matrix $W(\eta)^{-1} := \left(A^T, B^T\right) \Sigma(\eta)^{-1} \left(A^T, B^T\right)^T$, and assuming that it is invertible, the distribution above implies that

$$p(\theta) \propto N\left(\mu(\eta), W(\eta)\right) ,$$
$$\mu(\eta) = W(\eta)\left(A^T, B^T\right)\Sigma(\eta)^{-1}\begin{pmatrix} w \\ 0 \end{pmatrix} .$$

As for case *2)*, a statistically meaningful way to fuse the information items $p_1(\cdot)$ and $p_2(\cdot)$ in (6) without explicitly specifying their dependencies is the one proposed by [15]. Here, given the prior knowledge, we want to construct a prior $p(\theta)$ that *best* represents the *real* distribution of the model parameters, meaning that one aims at obtaining the best projection of the real distribution within the set of possible models implicitly defined by choosing a specific model structure [1]. The results from [15] show that a weighted product of p_1 and p_2 minimizes the expected Kullback-Liebler distance to this best possible prior. Following this principle and given the prior information p_1, p_2, the combined prior on θ would follow as

$$p\big(\theta | p_1, p_2\big) \propto p_1^{\alpha_1} \cdot p_2^{\alpha_2} \quad (7)$$

for opportune fixed weights[1] α_1 and α_2 that for simplicity of notation may here be assumed as 1. Note that the parameters α_i here simply scale the second moments of the priors p_i, thus we have in (7) a product of two Gaussians, and $p(\theta|p_1, p_2)$ is rendered Gaussian. Computing $p(\theta)$ in this way then reduces to case *1)* above though with $\Sigma_{12} = 0$, thus

$$W(\eta)^{-1} = A^T \Sigma_1^{-1}(\eta) A + B^T \Sigma_2^{-1}(\eta) B ,$$

and assuming that $W(\eta)^{-1}$ is full rank, it follows that

$$p\big(\theta | p_1, p_2\big) \sim N\left(W(\eta) A^T \Sigma_1^{-1} w, W(\eta)\right) .$$

The mean of this prior distribution on θ clearly depends on the hyperparameters that define $\Sigma_1(\eta)$ and $\Sigma_2(\eta)$.

In other words, in both cases *1)* and *2)* above, fusing an information source with nonzero first moment with some other hyperparametrized priors leads to a fused Gaussian prior whose mean is nonzero and hyperparametrized too.

A. Improper priors should be extended to proper ones

We illustrate here by a simple example why we sometimes may encounter improper priors that need to be extended to proper ones, and how this extension leads to prior means that depend on the hyperparameters as in the scenario above.

Example: Consider a stable linear time invariant SISO system that shall be modeled as an FIR system, for which we want to estimate its parameters (impulse response) θ, and for which we know approximately its steady state output value $\bar{y}$ for a given constant input $\bar{u}$. Such knowledge on a steady state pair $(\bar{u}, \bar{y})$ about a system is commonly exemplified as available prior knowledge [11], [20], [28]. This prior information connects with the estimand θ via the relationship

$$\bar{y} = \bar{u}\left(\sum_{i=1}^{k} \theta_i\right) = \bar{u}\mathbb{1} \cdot \theta \quad (8)$$

with $\mathbb{1}$ the (row)-vector of ones whose length is equal to the one of the vector θ, shall be treated as uncertain. One may then for simplicity impose a zero-mean Gaussian uncertainty on this relation, i.e., assume $(\bar{y} - \bar{u}\mathbb{1} \cdot \theta) \sim N(0, c)$, implying that there is some unknown additive error to (8), with some variance denoted c.

[1] We note that selecting these weights as proposed in [15] corresponds actually to consider them as hyperparameters, and estimating them via an empirical Bayes approach, and can even be included in η and tuned as in Section IV. We send the interested reader back to that paper for more information, and assume α_1 and α_2 in the remainder of this section as known, for simplicity.

As the previous example illustrates, prior information may be translated in a distribution on a *linear combination* of θ, i.e., of the form $p_1(A\theta) \sim N(w, c)$ (continuing with the example above, here $w = \bar{y}$, $A = \bar{u}\mathbb{1} \in \mathbb{R}^{1\times p}$ and c is the variance of this random variable). The matrix A may thus be singular, and actually this is always the case for any nontrivial dynamical model where the information is provided only on its steady state behavior. A singular A leads then to a precision matrix W^{-1} that is not invertible (to see this one may again take the example above, selecting $p(\theta) = p_1(A\theta)$), and thus to a specification of a prior $p(\theta)$ that is not a proper probability distribution. An improper prior on θ gives then an improper prior distribution for y, for which the marginal likelihood does not exist since it can not be normalized. It is therefore not possible to select a value for η through maximizing the marginal likelihood. Hence, in order to use such a method of selecting the hyperparameters, it is necessary to define some additional prior information p_2 that, in addition to p_1, makes the overall prior $p(\theta)$ proper. This can be accomplished as indicated above with using Σ_2 equal to, e.g., the TC or DC kernel [7] (if such prior knowledge exists) or even simply equal to λI (corresponding to selecting a ridge regression prior covariance, a prior that carries little information about the parameter correlations since embedding the belief that the parameters θ_i are i.i.d.). Equivalently, one may say that p_1 is added to extend p_2.

Remark 1 As explained, the maximum marginal likelihood is not available when the prior is improper, so other means of selecting the hyperparameters must be followed. Improper priors may yield proper posteriors, though while performing inference there are multiple pitfalls and paradoxes one may encounter with improper priors [9], [16], and should for these (and other) reasons generally be avoided.

IV. Improving the cost for numerically evaluating the marginal likelihood $p(y|\eta)$

We here show how it is possible to solve the marginal likelihood maximization problem starting from proper priors with nonzero means in a numerically more efficient way when $n \gg k$. In doing this we revisit and adapt some algebraic manipulations that are used in similar contexts to make the evaluation of similar cost functions so to have a leading order of $O(k^3)$ instead of $O(n^3)$ [24]. The matrix inversion lemma and Sylvester's determinant theorem are applied in ways that are essentially equivalent to those presented in [3], [4] and [6], and will be included for ease of readability and reproducibility of the results. We also use properties of QR decompositions, see [29], analogously to the application in [6]. We note again that the contribution in this section is generalizing the algorithm from [6] to allowing a nonzero Xm in (5), and where the generalized algorithm retains the order of complexity.

We assume that $\text{rank}(X,\, y) = k+1$, i.e. that y is not in the span of X. This assumption is present in all the mentioned related works in order to improve the efficiency of evaluating the marginal likelihood. If the rank is not full, it means that at least one θ from (1) is able to reproduce the output y perfectly, which is rarely the case, due to noise, unmodeled effects and a desire to obtain model structures that are "as simple as appropriate". If the assumptions from (1) indeed hold for X, y then the rank is almost surely full. Note that whether σ^2 is treated as unknown or not does not matter for the derivation of the objective functions below.

Assume then that the inverse of the prior covariance matrix is available[2] meaning that computing V^{-1} does not require matrix inversion, as it happens, e.g., for the DC and TC kernels, or in the case described in Section III. We start then from considering that the condition number of V may be very large for many useful kernels, especially those that encourage stability, and that this potentially leads to numerical instabilities. To circumvent this one may then exploit the knowledge of a closed form expression of V^{-1}, and its Cholesky decomposition $V^{-1} = DD^T$.

Our first goal is thus to rewrite (5) in terms of V^{-1} instead of V. To do so we may apply the determinant theorem to the $\log\det$ term in the right hand side of (5), resulting in

$$\begin{aligned}\log\det\left(\sigma^2 I_n(I_n + XVX^T\sigma^{-2})\right) &= n\log\sigma^2 \\ &\quad + \log\det(I_p + VX^TX\sigma^{-2}) \\ &= (n-p)\log\sigma^2 + \log\det V + \log\det(\sigma^2V^{-1} + X^TX).\end{aligned} \tag{9}$$

We may also apply the matrix inversion lemma to the other term in (5) involving V, to obtain

$$\begin{aligned}&\left(XVX^T + \sigma^2 I\right)^{-1} = \\ &\quad I_n\sigma^{-2} - X\left(\sigma^2V^{-1} + X^TX\right)^{-1}X^T\sigma^{-2}\,.\end{aligned} \tag{10}$$

By means of these two rewritings, the cost function (5) is transformed into something containing the term

$$\left(\sigma^2V^{-1} + X^TX\right)^{-1}X^T(y - Xm)\,. \tag{11}$$

As stated in [6], the computation of this term is prone to numerical errors if approached naively, and this may corrupt the accuracy of evaluating the cost function (5), which is crucial. This in practice follows from the fact that many popular prior kernels (e.g. TC, DC, SS) have a covariance matrix V which is numerically ill-conditioned due to very small matrix entries, or simply because X^TX itself is ill-conditioned. Therefore, solving (11) with the more accurate QR decomposition is highly useful to improve and stabilize the algorithm. Our second goal is thus to use the fact that $V^{-1} = DD^T$ to improve such numerical properties.

Firstly, one should realize that (11) is the solution to

$$\min_x \left\| \begin{pmatrix} X \\ \sigma D^T \end{pmatrix} x - \begin{pmatrix} y - Xm \\ 0 \end{pmatrix} \right\|^2 , \tag{12}$$

appointing us to consider the *thin* QR decomposition [29] of

$$\begin{pmatrix} X & y - Xm \\ \sigma D^T & 0 \end{pmatrix} = QR \in \mathbb{R}^{2n\times(p+1)} \tag{13}$$

[2] We recall that when V^{-1} is not easily available one is in a situation as in [6]. In this case one may rather rely on the Cholesky decomposition $V = LL^T$. In this case though the condition number of V may limit the region for which η leads to numerically stable solutions.

which is defined s.t. Q has dimension $2n \times (p+1)$ and R has dimension $(p+1) \times (p+1)$. By construction Q will moreover be orthogonal (orthogonal unit vector columns).

Since, by assumption, y is not in the span of X, then also $y - Xm(\eta)$ is not in the span of X for any η. Thus, $\text{rank}(X, y - Xm(\eta)) = p+1$ for any η. Then the left hand side matrix in (13) has full column rank. Theorem 5.2.2 of [29] then states that Q, R above is unique and the diagonal of R has positive entries. Then in the thin QR decomposition (13) we define $R_1 \in \mathbb{R}^{p\times p}, R_2 \in \mathbb{R}^p, r \in \mathbb{R}$ s.t.

$$R = \begin{pmatrix} R_1 & R_2 \\ 0 & r \end{pmatrix} . \tag{14}$$

Using the fact that $Q^TQ = I_{n+1}$ we get that

$$\sigma V^{-1} + X^TX = R_1^TR_1 \tag{15a}$$

$$X^T(y - Xm) = R_1^TR_2 \tag{15b}$$

$$(y - Xm)^T(y - Xm) = R_2^TR_2 + r^2 \tag{15c}$$

which combined with (10) simplifies the evaluation of the first term in (5) to

$$\begin{aligned}&(y - Xm)^T(I_n\sigma^2 + XVX^T)^{-1}(y - Xm) = \\ &(R_2^TR_2 + r^2)/\sigma^2 - R_2^TR_1(R_1^TR_1)^{-1}R_1^TR_2/\sigma^2 \\ &= r^2/\sigma^2 .\end{aligned} \tag{16}$$

Combining this with (9) yields that the cost function (5) in total can be rewritten to

$$r^2/\sigma^2 + (n-p)\log\sigma^2 + 2\log\det R_1 . \tag{17}$$

A. Further improvements exploiting QR factorizations

It turns out that the particular structure of the problem admits a way to speed up the QR factorization (13), which can be expensive, reducing it from dimension $2n\times(p+1)$ to $2p\times(p+1)$. This fact and the further improvements contained here have origin in [6], though generalized to $m(\eta) \neq 0$.

Consider the thin QR factorization

$$[X, y] = Q_d[R_{d1}, R_{d2}] . \tag{18}$$

Furthermore, consider the QR factorization of

$$\begin{pmatrix} R_{d1} & R_{d2} - R_{d1}m \\ \sigma D^T & 0 \end{pmatrix} = Q_cR_c \tag{19}$$

This is now a QR factorization of size $2p \times (p+1)$. Then from combining (13), (18) and (19) we have that

$$\begin{aligned}&\begin{pmatrix} Q_d & 0 \\ 0 & I_p \end{pmatrix}\begin{pmatrix} R_{d1} & R_{d2} - R_{d1}m \\ \sigma D^T & 0 \end{pmatrix} = \begin{pmatrix} X & y - Xm \\ \sigma D^T & 0 \end{pmatrix} \\ &= \begin{pmatrix} Q_d & 0 \\ 0 & I_p \end{pmatrix} Q_cR_c = QR\end{aligned} \tag{20}$$

By assumptions above giving uniqueness of the thin QR decomposition we have that $R = R_c$. Therefore, it suffices to solve (19) to obtain R. Therefore, computing R_{d1}, R_{d2} beforehand lets us solve a smaller problem (when $k < n$).

To evaluate the cost function one should (given R_{d1}, R_{d2}); first calculate $m(\eta), V(\eta)^{-1}$, before calculating D, then calculate $R = R_c$ from (19). Evaluation of the marginal likelihood is obtained as in (17). This approach with m being nonzero has only marginally higher computational complexity compared to $m = 0$, and same leading order of complexity. The generalization of the previous algorithms is seen in step (19), where the entry $R_{d2} - R_{d1}m$ has replaced R_{d2}. Therefore, only this step of the procedure has a computational load that is only linearly larger than the case when $m = 0$. The order of complexity is thus the same.

B. Computing the posterior

The insights about the QR factorization above yields an efficient way of computing the posterior parameter variance V^* and mean m^*. For some fixed η, using (3) and (15), we get that the posterior variance is

$$V^* = \sigma^2(R_1^TR_1)^{-1} ,$$

and that the posterior mean is

$$\begin{aligned}m^* &= V^*\left(X^Ty/\sigma^2 + V^{-1}m\right) \\ &= \left(R_1^TR_1\right)^{-1}\left(X^Ty \pm X^TXm + \sigma^2V^{-1}m\right) \\ &= \left(R_1^TR_1\right)^{-1}X^T(y - Xm) \\ &\quad + \left(R_1^TR_1\right)^{-1}(X^TX + \sigma^2V^{-1})m \\ &= (R_1)^{-1}R_2 + m ,\end{aligned}$$

For prior mean $m = 0$ all the above equations reduce to the procedures from *Algorithm 1* [3] and (with minor modifications) *Algorithm 2* [6].

Remark 2 The cost of evaluating the marginal likelihood can also be further improved by using the specific structure that many prior covariance matrices will have, also the ones here, though this is out of the scope of this article. See [3] for examples (there specifically for the DC kernel).

V. Simulation examples

We illustrate now how the above algorithm and ideas can be used to incorporate knowledge about the steady state for system identification purposes. The purpose of the simulations below is to show that, when the specific assumptions we posed above holds, it follows that • the new, extended empirical Bayes method proposed in this manuscript is a stable and efficient tool for selecting hyperparameters, • the proposed way to include steady state information is meaningful for the purpose of modeling dynamical systems, and • one can expect some improvement in fitting impulse responses for systems like those simulated here w.r.t. the algorithms cited in the introduction. As these works, we use the impulse response fit as a common metric for assessing the precision of the posterior mean m^* as an estimate of the true impulse response θ^*. Below we thus use the index

$$\text{fit} = 100 \cdot \left(1 - \frac{\sqrt{\sum_i(m_i^* - \theta_i^*)^2}}{\sqrt{\sum_i(\theta_i^*)^2}}\right) . \tag{21}$$

The simulations below assume then that we want to estimate an FIR model by means of data collected from a stable dynamical LTI system, and that some steady state estimate $(\bar{u}, \bar{y})$ is available. The example of Section III-A shows how to define a p_1 type of prior information, while some p_2 (see Section III) is needed since the steady state information is not enough for forming a proper prior.

For p_1 we let c denote the scalar variance and treat it as a hyperparameter to be tuned. For the distribution p_2 we choose the TC kernel, i.e.

$$\begin{aligned} p_2 &\sim N(0, K) \\ K_{i,j} &= \lambda \cdot \alpha^{\max(i,j)} \end{aligned} \tag{22}$$

where $\lambda \in \mathbb{R}^+, \alpha \in (0,1)$ are hyperparameters to be tuned, in addition to c. Together this gives a proper prior of the form (7), specifically

$$\begin{aligned} p(\theta | p_1, p_2) &\sim N(\Sigma \mathbb{1}^T c^{-1} \bar{u}\bar{y}\,,\, \Sigma) \\ \Sigma &:= \left(\bar{u}^2 c^{-1} \mathbb{1}^T \mathbb{1} + K\right)^{-1} \end{aligned} \tag{23}$$

Note that $\mathbb{1}^T\mathbb{1}$ is simply a square matrix with all entries 1. We then refer to modeling the system (23) with the proposed empirical Bayes procedure as *Method 1*, with hyperparameters λ, α, c.

Method 1 shall then be compared against *Method 2*, i.e. a similar approach that excludes the information p_1 from the prior, i.e. only using the TC kernel p_2. *Method 2* corresponds thus to the case where the variance c is $c \to \infty$. It thus only has two hyperparameters, λ, α.

Remark 3 As a disclaimer, note that the degree of improvement from *Method 2* to *Method 1* depends on the application; the actual system, number of inputs and outputs, model structure and data quality, will all influence the potential the extended prior has to improve (or disturb) the model. The goal is not to determine a specific degree of improvement, but rather see that the generalized algorithm works and that this way of including prior knowledge is meaningful. For some identification scenarios the steady state information will not be too valuable to improve the impulse response fit. Moreover, in other specific scenarios, applying other modeling approaches may also be more effective. We focus here on assessing the effect of excluding p_1 from the prior.

A. Implementation and evaluation aspects

We do 1000 iterations of the following simulation: *Step 1)* randomly construct a stable, discrete-time, minimum phase, linear system of order 14 with damping ratio in the range $(0.2, 1)$. *Step 2)* excite the system with a discrete white noise input of length 200. *Step 3)* add Gaussian white noise to the simulated output so to reach a SNR of 7. *Step 4)* compute the DC gain of the sampled system $\bar{y}$, corresponding to the constant input $\bar{u} = 1$. *Step 5)* maximize the marginal likelihood as described in the previous section, computing η^* that optimizes (5). *Step 6)* estimate the system with an FIR model of length $k = 70$ with η^* and (3).

The marginal likelihood maximization is done using standard, derivative-free optimization packages from the *scipy 1.11.3* Python package, and involves evaluating the marginal likelihood many times for each maximization. With $k = 70$ and $n = 200$ the here proposed way of evaluating the marginal likelihood is over 20 times faster than the more direct approach of (4). In many applications this factor will be much larger, and for MIMO systems the difference in computing time will be even more pronounced.

We compare *Method 1* and *Method 2* with the data as simulated above, with the following three tests:

Test 1: perform the steps above using the correct DC values of $\bar{u}, \bar{y}$,

Test 2: perform the same steps using though, instead of the correct DC value for $\bar{y}$, its value amplified by 10%,

Test 3: insert in step 5 an *oracle* that determines that optimal hyperparameters for maximizing the impulse response fit of the posterior mean, i.e. using (3) to maximize (21) (not implementable in practice).

The different methods will lead to different α, λ to define the TC kernel that shall be used. These differences in these values are then indicative of the different robustness properties of the two methods. For comparing the two methods we also compute the mean difference between the impulse response fits of Method 1 and Method 2, along with the standard deviation of this difference, denoted by M1 - M2.

B. Numerical results

The results of the three tests are summarized in Table I, presenting the mean fit and standard deviation of this mean over the 1000 simulated systems and signals, for Method 1, Method 2 and the difference M1 - M2.

		Mean fit	(Std. dev. of fit)
Test 1			
	Method 1	85.44	(0.218)
	Method 2	84.53	(0.315)
	M1 - M2	0.92	(0.216)
Test 2			
	Method 1	84.64	(0.334)
	Method 2	84.25	(0.343)
	M1 - M2	0.39	(0.241)
Test 3			
	Method 1	88.51	(0.410)
	Method 2	88.22	(0.421)
	M1 - M2	0.29	(0.134)

TABLE I

COMPARISON OF THE FIT LEVELS OBTAINABLE USING METHOD 1 (THE PROPOSED ONE) AND METHOD 2 (FOR WHICH p_1 IS EXCLUDED FROM THE PRIOR) ON THE DIFFERENT TESTS, I.E. TEST 1 (USING THE CORRECT DC VALUES), TEST 2 (USING PERTURBED VALUES FOR THE DC GAIN) AND TEST 3 (USING AN ORACLE).

We note that there exist simulations where Method 2 performs seemingly better than Method 1 (also for Test 3). E.g., for oscillatory impulse responses the DC gain is less descriptive of the system properties, and so in practice the information from p_1 can then disturb the optimization. A more thorough analytical explanation of this will be pursued in future works. On *average* though Method 1 obtains a small improvement over Method 2, that is statistically significant for various tests, and that certify the usefulness of using p_1 as a source of information in the prior. Furthermore, we can note that the advantage of Method 1 generally increases with poorer data quality (changing the quantity, noise and excitation), something that is expected since Method 1 has broader regularization possibilities.

From inspecting the numerical properties of the optimization schemes it is clear that the Method 1 comes with a cost function that requires more iterations to arrive at a similar optimization level than for Method 2, which can be due to the extra hyperparameter c and/or the fact that it becomes more nonlinear from the introduction of the extra penalty, and is therefore expected. These extra evaluations are also observed in the tests. Using Jacobians and/or Hessians in the optimization may thus turn out to be especially important when embedding multiple sources of prior information like done here, and investigating this is left for future works. This though seems like an insignificant price to pay in order to include more prior knowledge and improve the model.

The tests confirm the intuition for which including accurate steady state information $\bar{u}, \bar{y}$ maximizes the increase of performance of Method 1. Indeed as seen in test 2), perturbing $\bar{y}$ makes the difference between the methods less pronounced. We also observe (not included here) that when perturbing $\bar{y}$ by as much as 20% the difference M1 - M2 is still positive but with a standard deviation that can not imply any statistical significance. We note again that the results are specific to the examples simulated here. Test 3) indicates that optimally the advantage of Method 1 over Method 2 is indeed moderate but significant, meaning that including the information in p_1 indeed improves the modeling. Test 3) indicates moreover that the relative improvements seen from Test 1) and Test 2) are reasonable.

VI. Conclusion

Tuning of hyperparameters in regularization-based models is crucial and can be a major crux, either in terms of numerical complexity or instability due to ill-conditioned matrices. We have presented an extended algorithm that allows for efficient marginal likelihood maximization wrt. hyperparameters in priors where the prior mean depends on the hyperparameters themselves. The new method allows to extend existing regularization based methods to the non-null mean case while retaining the computational complexity and stability of the whole algorithmic structure. The capabilities of the new method were illustrated on simulation examples, showing that including prior information on the steady state in this way can be beneficial. The algorithm finds meaningful values for the hyperparameters, and the resulting models generally obtain an improved impulse response fit.

As future works we wish to test out the algorithm on real-life examples to see if other system models with nonzero prior mean can yield good posterior estimates. This empirical Bayes optimization will likely benefit particularly from using the Jacobian and Hessian of the cost function. Lastly, we wish to incorporate more extensive prior information, using the empirical Bayes approach presented here.

References

[1] L. Berec and M. Kárný. Identification of reality in bayesian context. In *Computer Intensive Methods in Control and Signal Processing: The Curse of Dimensionality*. Birkhäuser Boston, 1997.

[2] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[3] F. P. Carli, T. Chen, and L. Ljung. Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, 62(3):1471–1477, 2016.

[4] F. P. Carli, A. Chiuso, and G. Pillonetto. Efficient algorithms for large scale linear system identification using stable spline estimators. *IFAC Proceedings Volumes*, 45(16):119–124, 2012.

[5] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, 59(11):2933–2945, 2014.

[6] T. Chen and L. Ljung. Implementation of algorithms for tuning parameters in regularized least squares problems in system identification. *Automatica*, 49(7):2213–2220, 2013.

[7] T. Chen and L. Ljung. What can regularization offer for estimation of dynamical systems? *IFAC Proceedings Volumes*, 46(11):1–8, 2013.

[8] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and gaussian processes—revisited. *Automatica*, 48(8):1525–1535, 2012.

[9] A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization paradoxes in bayesian and structural inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 35(2):189–213, 1973.

[10] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge Univ. Press, 2012.

[11] Y. Fujimoto and T. Sugie. Kernel-based impulse response estimation with a priori knowledge on the DC gain. *IEEE Control Systems Letters*, 2(4):713–718, 2018.

[12] A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 2017.

[13] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[14] M. Kárný. Quantification of prior knowledge about global characteristics of linear normal model. *Kybernetika*, 20(5):376–385, 1984.

[15] M. Kárný, N. Khailova, P. Nedoma, and J. Böhm. Quantification of prior information revised. *International Journal of Adaptive Control and Signal Processing*, 15(1):65–84, 2001.

[16] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[17] L. Ljung. *System identification : theory for the user*. Prentice Hall information and system sciences series. Prentice Hall PTR, Upper Saddle River, N.J, 2nd ed. edition, 1999.

[18] D. J. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

[19] A. Marconato, M. Schoukens, and J. Schoukens. Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11(2):194–204, 2017.

[20] T. Peter and O. Nelles. Gray-box regularized fir modeling for linear system identification. In *Proceedings, 28th Workshop on Computational Intelligence*, pages 113–128, 2018.

[21] V. Peterka. Bayesian approach to system identification. In *Trends and Progress in System identification*, pages 239–304. Elsevier, 1981.

[22] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung. *Regularized system identification: Learning dynamic models from data*. Springer Nature, 2022.

[23] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3):657–682, 2014.

[24] C. E. Rasmussen, C. K. Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

[25] C. S. Reese, A. G. Wilson, M. Hamada, H. F. Martz, and K. J. Ryan. Integrated analysis of computer and physical experiments. *Technometrics*, 46(2):153–164, 2004.

[26] C. Roberts. *The Bayesian Choice*. Speinger Verlag New York, 2007.

[27] B. Sohlberg and E. W. Jacobsen. Grey box modelling–branches and experiences. *IFAC Proceedings Volumes*, 41(2):11415–11420, 2008.

[28] P. Trnka and V. Havlena. Subspace like identification incorporating prior information. *Automatica*, 45(4):1086–1091, 2009.

[29] C. F. Van Loan and G. Golub. Matrix computations (Johns Hopkins studies in mathematical sciences). *Matrix Computations*, 5, 1996.